Sparse long blocks and the variance of the LCS

S. Amsalu* C. Houdré† H. Matzinger‡

March 2, 2013

Abstract

Consider two random strings having the same length and generated by two mutually independent iid sequences taking values uniformly in a common finite alphabet. We study the order of the variance of the longest common subsequence (LCS) of these strings when long blocks, or other types of atypical substrings, are sparsely added into one of them. We show that the existence of the derivative of the mean LCS-curve at its maximum implies that this order is linear in the length of the strings. We also argue that our proofs carry over to many models used by computational biologists to simulate DNA-sequences.

1 Introduction

Let x and y be two finite strings. A common subsequence of x and y is a subsequence which is a subsequence of both x and y, while a longest common subsequence (LCS) is a common subsequence of maximal length. For example, let $x = heinrich$ and let $y = enerico$. Then $z = ni$ is a common subsequence of x and y, indicating that the string ni can be obtained from both x and y by just deleting letters. Common subsequences can be represented via alignments, and for this the letters which are part of the subsequence get aligned while the other letters get aligned with gaps. The subsequence ni, corresponds to the alignment with gaps:

x	h		e	i	n		r		i	c			h
y		e			n	e		r	i		c	o	

The common subsequence ni is not of maximal length, the LCS is $enric$, and the corresponding alignment is given by:

x		h	e	i	n		r	i	c		h
y			e		n	e	r	i	c	o	

*Department of Information Science, Faculty of Informatics, University of Addis Ababa, Ethiopia

†School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332
houdre@math.gatech.edu. Supported in part by NSA Grant H98230-09-1-0017.

‡School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332
matzi@math.gatech.edu. Supported in part by NSA Grant H98230-09-1-0017.

Often, a long LCS indicates that the strings are related. In this article, we only consider alignments which align same letter pairs, every such alignment defines a common subsequence, and the length of the subsequence corresponding to an alignment is called the *score of the alignment*. The alignment representing a LCS is also called an *optimal alignment* (OA). In the above example, the length of the LCS is five which is denoted by:

$$|LCS(heinrich, enerico)| = |LCS(x, y)| = 5.$$

Longest Common Subsequences (LCS) and Optimal Alignments (OA) are important tools used for string matching in Computational Biology and Computational Linguistics [8, 24, 25]. A main application is to the automatic recognition of related DNA pieces. In that context, it is anticipated that if two DNA-strings have a common ancestor, then they will have a long LCS. Could it be that by chance (bad luck), unrelated (independent) strings have nonetheless a long LCS? How likely is such an event? This, of course, depends on the probabilistic model generating the strings. To answer the previous questions, the behavior, for n large, of both the expectation $\mathbb{E}LC_n$ and the variance $\operatorname{Var} LC_n$ need to be understood. (Throughout LC_n is the length of the LCS of the random strings $X = X_1 \cdots X_n$ and $Y = Y_1 \cdots Y_n$.)

The asymptotic behavior of the expectation and the variance of the length of the LCS of two independent random strings has been studied by probabilists, physicists, computer scientists and computational biologists. It can also be formulated as a last passage percolation problem with dependent weights. The problem of finding the fluctuation order for first and last passage percolation has been open for a while. There has been, however, a well-known breakthrough for a related problem, that is for the Longest Increasing Subsequence (LIS) of a random permutation [5] or of a random word [16, 17, 23]. For the LIS of a random permutation, the order of the fluctuation is the cubic root of the expectation and not its square root. For the LCS case, the expectation is of order n, and so if the fluctuations were also a cubic root of the expectation, then $\operatorname{Var} LC_n$ should be of order $n^{2/3}$. This is the order of magnitude conjectured in [9] for which several heuristic proofs have been claimed. This conjectured order might even seem more plausible in view of the recent solution [19, 20] to the Bernoulli matching problem where the $n^{2/3}$ order is shown to be correct. We believe this order to be incorrect for the LCS. However, for short sequences this order might be what one approximately observes in simulations. For the LCS-problem, in case of independent iid strings the order of magnitude of the variance is, in general, not known. (Except for various cases of binary sequences, see [14], [15], [18], in which case the variance is asymptotically linear in the length of the strings considered.) In the present article, we determine the correct asymptotic order of the variance for the LCS of uniform iid sequences "with artificially added impurities" provided the mean LCS-curve (see (1.5)) is differential at its maximum.

The cubic root claims might happen because the LCS length can be viewed as a last passage percolation problem with dependent weights, and so, for short sequences the dependence in the weights does not have a strong influence, the LCS then behaves as if the weights were independent. Let us now explain how the LCS-problem can be viewed as a last passage percolation problem:

Let the set of vertices be

$$V := \{0, 1, 2, \ldots, n\} \times \{0, 1, 2, \ldots, n\},$$

and let the set of oriented edges $\mathcal{E} \subset V \times V$ contain horizontal, vertical and diagonal edges. The horizontal edges are oriented to the right, while the vertical edges are oriented upwards, both having unit length. The diagonal edges point up-right at a 45-degree angle and have length $\sqrt{2}$. Hence,

$$\mathcal{E} := \{(v, v + e_1), (v + e_2, v), (v, v + e_3) : v \in V\},$$

where $e_1 := (1, 0)$, $e_2 := (0, 1)$ and $e_3 := (1, 1)$. With the horizontal and vertical edges, we associate a weight of 0. With the diagonal edge from (i, j) to $(i+1, j+1)$ we associate the weight 1 if $X_{i+1} = Y_{j+1}$ and $-\infty$ otherwise. In this manner, we obtain that $LC_n := |LCS(X_1X_2 \ldots X_n; Y_1Y_2 \ldots Y_n)|$, is equal to the total weight of the heaviest path going from $(0, 0)$ to (n, n). Note that the weights in our 2-dimensional graph are not "truly 2-dimensional" and they depend only on the one dimensional sequences $X = X_1 \ldots X_n$ and $Y = Y_1 \ldots Y_n$. In our opinion, this is the reason for the order of magnitude of the variance of the LCS to be different from other first/last passage-related models.

A subadditivity argument pioneered in [9] shows that the existence of

$$\gamma_k^* := \lim_{n\to\infty} \frac{\mathbb{E}\, LC_n}{n}, \tag{1.1}$$

where X and Y are two stationary ergodic strings independent of each other and where the constant $\gamma_k^* > 0$ depends on the distribution of X and Y and on the size k of the alphabet. Even for the simplest distributions, such as iid strings with binary equiprobable letters, the exact value of γ_k^* is unknown, and extensive simulations have been performed to obtain approximate values [4, 7, 10, 11, 12, 13].

The speed of convergence to the expected length in (1.1) was further determined in [1, 2], showing that for iid sequences,

$$\gamma_k^* n - C\sqrt{n \log n} \le \mathbb{E}\, LC_n \le \gamma_k^* n, \tag{1.2}$$

where $C > 0$ is a constant depending neither on n nor on the distributon of X_1.

As previously mentioned, there exist contradicting conjectures for the order of the variance of the LCS. Our present result (Theorem 2.1) establishes the order conjectured in [26] for an iid distribution. We prove it, however, for iid sequences with added impurities (sparse long blocks or atypical substrings), assuming also that the mean LCS-curve has a well-defined derivative at its maximum, a condition which has not been proved to hold in full generality so far. The mean LCS-curve is the rescaled expectation of the LCS when the two sequences are taken to be of different length but in a fixed proportion. (See (1.5).) The impurities or "long blocks" as we call them, are substrings consisting only of one symbol which can be different from block to block. For that model, the variance is shown to be of order $\Theta(n)$, i.e., there exist two constants $C_2 > C_1 > 0$ independent of

n, such that $C_1 n \leq \operatorname{Var} LC_n \leq C_2 n$, for all natural number n. (Here LC_n is the length of the LCS of the two independent sequences X and Y of length n. One of the two sequences is to have sparsely added long blocks.) It is rather interesting that the mere differentiability of the mean curve at its maximum, implies a certain order of magnitude for the variance. Note that [21, 22] proved that $\operatorname{Var} LC_n \leq n$, and so only good lower bounds for the variance of LC_n are needed. (Simulation studies are not that numerous in case of the variance and at times contradict each other.)

Still for iid sequences with k equiprobable letters, the order of $\operatorname{Var} LC_n$ remains unknown. We hope nonetheless, that similar ideas could be helpful in fully tackling this problem.

Overview of the main result of this paper

We first need a few definitions: Let $V_1, V_2, \ldots$ and $W_1, W_2, \ldots$ be two independent iid sequences with k equiprobable letters, and let

$$\gamma_k^* := \lim_{n\to\infty} \frac{\mathbb{E}|LCS(V_1 V_2 \ldots V_n; W_1 W_2 \ldots W_n)|}{n}.$$

As already mentioned, the exact value of γ_k^* is unknown, but lower and upper bounds are available, e.g.,

k	2	3	4	$\cdots$
γ_k^*	0.812	0.717	0.654	$\cdots$

(1.3)

where the precision in the above table is about ± 0.01. The expected length of the LCS of two independent iid sequences both of length n is thus about $\gamma_k^* n$, up to an error term of order not more than a constant times $\sqrt{n \log n}$ (see (1.2)).

We can also consider two sequences of different lengths, but in such a way that the two lengths are in a fixed proportion of each other. To do so, let

$$\gamma_k(n, q) := \frac{\mathbb{E}|LCS(V_1 V_2 V_3 \ldots V_{n-nq}; W_1 W_2 \ldots W_{n+nq})|}{n}, \tag{1.4}$$

where $q \in [-1, 1]$, and let

$$\gamma_k(q) := \lim_{n\to\infty} \gamma_k(n, q), \tag{1.5}$$

which again exists by subadditivity arguments. The function $q \mapsto \gamma_k(q)$ is called the *mean LCS-function*, it is symmetric around $q = 0$ and concave and it thus has a maximum at $q = 0$ which is equal to γ_k^*. This function corresponds to the wet-region-shape in first passage percolation.

The fluctuation result (Theorem 2.1) of the present paper shows that the mere existence of the derivative, at its maximum, of the function $q \mapsto \gamma_k(q)$ implies that

$$\operatorname{Var} LC_n = \Theta(n),$$

for a model with sparse long blocks added into an iid sequence with k-equiprobable letters. (A *block* is a maximal contiguous substring consisting of only one symbol.) The model

considered will be described in detail at the beginning of Section 2 but let us, nevertheless, already give an overview of it. Let β and p to be reals (independent of n) such that

$$\frac{1}{2} < \beta < 1$$

and

$$0 < p < 1.$$

Then, take d large, but fixed, while n goes to infinity (the choice of the size of d depends on the choice of β, but $n = 2dm$). Now add into the sequence X long blocks of length about $\ell = d^{\beta}$. The possible locations for the long blocks are, say, $d, 3d, 5d, \ldots, 2dm - d$, where again $n = 2dm$. For each possible location throw independently the same (possibly biased) coin to decide whether or not to place a long block there, and the probability to place in a given location a long block is p. The sequence X and Y both have length n, and the string Y is iid with k-equiprobable letters. Moreover, X is iid with the same k equiprobable letters except possibly in the places where we have long blocks and letters could be different from long blocks to long blocks.

Let us present an example. Take $d = 5$ and $\ell = 4$, while $m = 2$. The length of the sequences X and Y is thus $n = 2dm = 20$. Consider binary sequences so that $k = 2$. We are thus throwing an unbiased coin independently 20 times to obtain the sequence Y. For example we could have:

$$Y = 00101110100011010101.$$

Then, we throw our unbiased coin again 20 times to obtain the sequence X^*. The sequence X^* is thus also uniform iid, and we proceed to add long blocks into X^* in order to obtain the string X. The potential places for long blocks are the integer intervals $[d(2i-1) - \ell/2, d(2i-1) + \ell/2]$, where $i = 1, 2, \ldots, m$. In the present example, there are only $m = 2$ such intervals:

$$[3, 7] \text{ and } [13, 17]. \tag{1.6}$$

Assume that after having thrown our unbiased coin $n = 20$ times, we obtained for X^* the sequence

$$X^* = 01\mathbf{01010}00101\mathbf{00101}110,$$

where the bold face substrings could get replaced by long blocks. The next step is to throw a (possibly biased) coin for each of the intervals which could get a long block. (We will thus throw the coin $m = 2$ times.) In this way, we decide for each of the intervals in (1.6) whether or not there will be a long block covering it. If the corresponding Bernoulli random variable Z_i is equal to 1, then there will be a long block covering the interval $[d(2i-1) - \ell/2, d(2i-1) + \ell/2]$ and if it is equal to 0 then otherwise. The probability of a long block is thus equal to $\mathbb{P}(Z_i = 1) = p$. In the present example assume we throw our coin twice and obtain $Z_1 = 1$ and $Z_2 = 0$, then in the interval $[3, 7]$ we place a long block, say made up of zeros, while in $[13, 17]$ we leave things as they are. With these modifications we obtain:

$$X = 01\mathbf{00000}00101\mathbf{00101}110.$$

In other words, to obtain X from X^*, we simply fill each integer interval

$$[d(2i-1)-\ell/2, d(2i-1)+\ell/2]$$

for which $Z_i = 1$, with all the same bits and leave everything else unchanged. In our example,

X^*		0	1	**0**	**1**	**0**	**1**	**0**	0	0	1	0	1	**0**	**0**	**1**	**0**	**1**	1	1	0
X		0	1	**0**	**0**	**0**	**0**	**0**	0	0	1	0	1	**0**	**0**	**1**	**0**	**1**	1	1	0

.

As already mentioned, the main result of the present paper (Theorem 2.1) is that the order of magnitude of $\operatorname{Var} LC_n$ is linear, i.e.,

$$\operatorname{Var} LC_n = \Theta(n),$$

if the mean LCS curve γ_k is differentiable at its maximum γ_k^*. To prove this result, we take the order of the length of the inserted long blocks larger than $\sqrt{d}$, but smaller than d. More precisely, we take d and $p \in (0,1)$ fixed, while n the common length of X and Y goes to infinity. We take a parameter β not depending on d such that $1/2 < \beta < 1$, and set the length of the long blocks to be $\ell = d^\beta$. Our result holds, for all d large enough, but fixed, and assuming a block-length of $\ell = d^\beta$ (d does not need to be very large for our fluctuation result to hold). Note also that the length of the long block does not need to be exactly ℓ, it could be a little bigger, but this is of no real importance in the present investigation.

The present paper was partly motivated by remarks from computational biologists to the effect that DNA distribution is not homogeneous. Rather there are different parts, with different biological functions (exon, coding parts, non-coding parts, ...). These different parts, having different lengths and each having its own distribution, are often modeled by computational biologists using hidden Markov chains; the hidden states determining the parts. Once the hidden states are determined, the DNA-sequence is drawn, by using the corresponding distribution for each part.

The reader might wonder how realistic our present long block model is, in view of this hidden-Markov model. Why did we add long blocks in predetermined positions and why do they only get added into one sequence and not both? Also, in DNA-sequences there are typically no long blocks. Let us present the various restrictions of our model and explain which features are present only to simplify the already involved notation, but do not represent a fundamental restriction:

1. The first restriction is that we add long blocks in predetermined locations. This restriction is only there to simplify notation. The same proof works if we use a Poisson point process with intensity-parameter $\lambda = 1/2d$ to determine the locations of the long blocks. Also, we could take the length of the long blocks to be a geometric random variable with expectation $\ell = d^\beta$.

2. Another quite unnatural restriction is to put the long blocks only into one sequence. This is done again to simplify notations. If the starting location of the long blocks

is given by a Poisson point process with intensity $\lambda = 1/2d$, then we can add long blocks in both sequences X and Y. We would then use independent Poisson point processes with the same intensity for both X and Y. The proofs presented here work as well for this case.

3. The model with long blocks added in Poisson locations, is very similar to a 2-state hidden Markov chain. For this we could take the hidden states to be L and R. The state L would correspond to a long block, while R would be the places where the string is iid. The transition probabilities from L to R would be $1/d^\beta$, while from R to L it would be $1/2d$. Again, for this hidden 2-state model proofs very similar to the ones presented here will give the linear order of the variance, but the notations would have to become even more cumbersome.

4. In DNA-sequence, there are no long strings consisting only of one symbol. So, the long block model may at first not look very realistic. However, in place of long blocks, we can take pieces generated by another distribution. For this we take two ergodic distributions. Typically one could use a finite Markov chain or a hidden Markov model with finitely many hidden states. For each different part, we would use the corresponding distribution. We could use a hidden Markov model first to determine which positions belong to which part and then fill the part with strings obtained from the corresponding distribution. (These corresponding distributions will again typically be hidden Markov with finitely many hidden states or Markov or finite Markov, maybe even Gibbs.) The hidden states could again be L and R. (To simplify things here we assume that there are only two DNA-parts.) But this time the state L would not correspond to a long block. Rather we would have two stationary, ergodic distributions μ_L and μ_R. The places with hidden state R would get the DNA-sequence drawn using μ_R, while for the positions with hidden state L, we would draw the DNA-sequence from μ_L. The transition probabilities between L and R would be as before: from L to R it would be $1/d^\beta$, while from R to L it would be $1/2d$. We believe that our current approach to determine the order of the variance could work for this hidden Markov chain case, provided we had:

$$\gamma_{L,R}(q) < \gamma_R(q), \tag{1.7}$$

for all q in an appropriate closed interval around 0. Here, $\gamma_{L,R}(q)$ is the coefficient for the mixed model:

$$\gamma_{L,R}(q) := \lim_{n\to\infty} \frac{\mathbb{E}|LCS(V_1V_2V_3\dots V_{n-nq}; W_1W_2\dots W_{n+nq})|}{n},$$

when the string $V_1V_2\dots$ is drawn according to μ_L and $W_1W_2\dots$ is drawn according to μ_R. Similarly, $\gamma_R(q)$ is the parameter when both sequences are drawn according to μ_R:

$$\gamma_R(q) := \lim_{n\to\infty} \frac{\mathbb{E}|LCS(U_1U_2U_3\dots U_{n-nq}; W_1W_2\dots W_{n+nq})|}{n},$$

where both strings $U_1U_2\ldots$ and $W_1W_2\ldots$ are drawn independently from each other with distribution μ_R. The condition (1.7) makes sense: it seems clear that when we align two sequences drawn from the same distribution typically we should get a longer LCS than if we align sequences from a different distribution. This might be difficult to prove theoretically. (This is why we consider long blocks, since they make this kind of condition easily verifiable.) Also, we would need for γ_R to have a well defined derivative at all its maximal points. In fact, instead of long blocks, any atypical long substrings such that its asymptotic expected LCS is smaller than γ_k^* will do.

5. A true restriction of our method is that the long blocks are of order greater than $\sqrt{d}$. Our current methodology does not carry over when this is lacking. It should be noted however, that different parts (exon, coding, non-coding part) of DNA are often pretty long, so the current assumption might not be totally unrealistic. We do not know how to treat the case, when the added long blocks have length below $\sqrt{d}$.

6. We also assumed that the long blocks have length of order below a constant time d. For the cases with long blocks of order d times a constant, a different paper would need to be written. But this seems clearly within reach, considering the present results.

Summarizing: if we are willing to accept that the function γ_R has a well defined derivative at its maxima and that the condition (1.7) holds, we probably should be able to get

$$\operatorname{Var} LC_n = \Theta(n),$$

for a whole range of distributions used in practice to model DNA. We plan to investigate this problem in the future.

Let us briefly describe the content of the rest of the paper. In the next section, the problem of the order of the variance is first reduced to the biased effect of long blocks. For this, we choose one long block at random and change it back to iid. Theorem 2.2 then states that if such a random alteration has typically a sufficiently strong biased effect, then the linear order of the variance follows. After Theorem 2.2, the rest of the section is devoted to establishing the biased effect of the long block replacement. It is shown that from the biased effect for a one-long-block situation (which was established in [3]), a biased effect follows for the many-long-blocks case.

2 Long blocks and the variance

In the present section, we consider strings of length n with many long blocks added. This many-long-blocks model was briefly explained in the introduction and it is precisely defined now: First, the string $Y = Y_1 \cdots Y_n$ is iid with k equiprobable symbols. Next, d is taken large but fixed as n goes to infinity and we partition the iid sequence $X^* = X_1^* \ldots X_n^*$

into pieces of length $2d$, so that $n = 2dm$. We then insert, say, in the middle of each of these pieces at most one long block, deciding at random which pieces get a long block of length ℓ and which do not. In all places where there is no long block, the sequence is iid with k equiprobable symbols. Let us explain in more details how this string X is defined: For each $i = 1, 2, \ldots, m$, let J_i be the interval

$$J_i := \left[(2i-1)d - \frac{\ell}{2}, (2i-1)d + \frac{\ell}{2}\right],$$

i.e., J_i is the i-th place where a long block could be introduced. We assume that $X^* = X_1^* X_2^* X_3^* \ldots X_n^*$ is iid uniform. The string X is equal to X^* everywhere except possibly at the places where we put long blocks:

$$X_i := X_i^*\,,\ \forall i \in [1, n] - \bigcup_{j=1}^{m} J_j.$$

Let Z_i be the Bernoulli random variable which, when equal to one, places a long block into the interval J_i. Hence,

$$Z_i := 1 \ \text{ implies } \ X_{j_1} = X_{j_2} \ \ \forall j_1, j_2 \in J_i,$$

and let $Z_1, Z_2, \ldots, Z_m$ be iid Bernoulli random variables with

$$\mathbb{P}(Z_i = 1) = p,$$

where $p \in (0, 1)$. Hence, p is nothing but the probability to have a long block introduced artificially into one of the possible locations. (The variables Z_1, Z_2,...,Z_m are all independent of X^* and Y, and the string $Y = Y_1 Y_2 \ldots Y_n$ is independent of X and X^*.) Moreover the strings are drawn from an alphabet $\mathcal{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_k\}$ with k equiprobable symbols:

$$\mathbb{P}(X_i = \alpha_j) = \mathbb{P}(X_i^* = \alpha_j) = \mathbb{P}(Y_i = \alpha_j) = \frac{1}{k},$$

for all $i = 1, \ldots, n = 2dm$ and all $j \in \{1, 2, \ldots, k\}$.

We are now ready to state the main result of the paper. It gives the asymptotic order of the variance of the LCS of X and Y for the distribution with many long blocks added. It is valid for any alphabet size k.

Theorem 2.1 *Let X and Y be two independent strings of length $n = 2dm$ drawn from an alphabet with k equiprobable letters, $k \geq 2$. Let Y be iid and let X be a string with artificially long blocks randomly placed into some of the location $J_1, J_2, \ldots, J_m$. Each of the locations has a probability p to receive a long block independently of the others. Outside the long block areas, the string X is iid. Let the concave function γ_k be differentiable at its maximum, then there exists d_1 such that for all $d \geq d_1$ independent of n,*

$$\operatorname{Var} LC_n = \Theta(n).$$

(Recall that d is held fixed while n goes to infinity.)

For the above theorem to hold, it is enough to show that the change of one long block (picked at random) induces an expected increase in the LCS. For this we choose in X one of the long blocks at random and change it back into iid. We assume that all the long blocks have equal probability to get chosen and the string obtained by changing one long block into iid is denoted by $\tilde{X} = \tilde{X}_1\tilde{X}_2\dots\tilde{X}_n$. Let us describe $\tilde{X}$ a little bit more formally: First recall that X^* denotes "the iid string X before the long blocks are introduced." Let N be the total number of long blocks in X, i.e.,

$$N := \sum_{i=1}^{m} Z_i,$$

and let $i(j)$ be the index of the j-th long block, i.e., if

$$\sum_{s=1}^{i-1} Z_s = j-1, \quad \sum_{s=1}^{i} Z_s = j,$$

then $i(j) := i$.

Next, let M be a random variable which is uniform on $\{1, 2, \dots, r\}$ when $N = r$:

$$\mathbb{P}(M = j \mid N = r) = \frac{1}{r}, \quad \forall j \le r,$$

and assume that conditionally on $N = r$, the variable M is independent of X, X^* and Y. The block we change has index $i(M)$, therefore

$$\mathbb{P}(\tilde{X}_s = X_s, \forall\, s \notin J_{i(M)}) = 1$$

and

$$\mathbb{P}(\tilde{X}_s = X^*_s, \forall\, s \in J_{i(M)}) = 1.$$

In other words, the strings X and $\tilde{X}$ are the same everywhere except on the interval $J_{i(M)}$, on that interval, $\tilde{X}$ is equal to the iid sequence X^*. We are now ready to formulate the result stating that in order to show that $\operatorname{Var} LC_n = \Theta(n)$, it is enough to prove that the randomly changed block typically has a positive biased effect on the length of the LCS:

Theorem 2.2 *If there exist two constants $c_1, c_2 > 0$ independent of n or d such that*

$$\mathbb{P}\left(\mathbb{E}\left(|LCS(\tilde{X};Y)| - |LCS(X;Y)|\,\middle|\, X, Y\right) \ge c_1 d^{\beta}\right) \ge 1 - e^{-c_2 n}, \tag{2.1}$$

for all d large enough (but independent of n), then

$$\operatorname{Var} LC_n = \Theta(n).$$

Proof. The idea is to represent $LC_n = |LCS(X;Y)|$ as a function f of a binomial random variable N with f linearly increasing along some scales. Note that if a function $f : \mathbb{R} \to \mathbb{R}$, is such that $f' \geq c$ and if T is a random variable with finite variance, then

$$\operatorname{Var} f(T) \geq c^2 \operatorname{Var} T. \tag{2.2}$$

Therefore, if N is a binomial random variable with parameters $m = n/2d$ and p,

$$\operatorname{Var} f(N) \geq \frac{c^2 np(1-p)}{2d}. \tag{2.3}$$

This last inequality gives the desired order for the variance of $f(N)$, i.e., $\operatorname{Var} f(N) = \Theta(n)$, and it remains to find a way to represent LC_n as $f(N)$, where f is a function which typically increases linearly.

This is done as follows: let $X(\ell)$ denote a string of length n whose distribution is the distribution of X conditional on the number of long blocks to be ℓ:

$$\mathcal{L}(X(\ell)) = \mathcal{L}(X \mid N = \ell),$$

where $\mathcal{L}$ stands for the law of the corresponding random variables. The strings $X(\ell)$ are all taken independent of Y and of N. We first simulate $X(m)$. For this, $X(m)$ is a string of length n with long blocks in every interval J_i for $i = 1, 2, \ldots, m$ and iid outside those intervals. Hence $X(m)$ is the string "with maximum number of long blocks inserted." Then we obtain $X(m-1)$ by choosing in $X(m)$ one long block at random and turning it into iid and proceeding by induction, once $X(\ell)$ is defined we obtain $X(\ell-1)$ by choosing one long block in $X(\ell)$ at random and turning it into iid. Again, all long blocks have same probability to get chosen. (We consider only the artificially inserted long blocks, and not blocks in the iid part which might be long by chance.) Next, let

$$LC_n(\ell) := |LCS(X(\ell); Y)|.$$

It is easy to notice that, with this construction, $X(\ell)$ has the same distribution as X conditional on $N = \ell$ and therefore that $X(N)$ has the same distribution as X. So LC_n has the same distribution as $LC_n(N)$ and

$$\operatorname{Var} LC_n = \operatorname{Var} LC_n(N).$$

Take now $f : \ell \mapsto LC_n(\ell)$. Note that by condition (2.1), $\ell \mapsto LC_n(\ell)$ behaves like a biased random walk path which insures that the function $LC_n(\cdot)$ tends to increase linearly. Clearly, $\ell \mapsto LC_n(\ell)$ is typically not going to increase at every step but rather on a $\log n$ scale. This is enough to get an inequality like (2.3), by extending techniques as in [6]. ∎

So far, we have reduced the problem to the biased effect of our random change. Next we need to prove that changing a randomly chosen long block into iid has the desired biased effect. This biased effect is a consequence of [3], establishing, in Theorem 2.1 and 2.2 there, inequalities such as (2.1) but for strings of length $2d$ with only one long block.

In other words, the probability in the one long block setting, not to have an increase linear in the length of the block is extremely unlikely as soon as d is not too small, but still fixed. With this result for one long block, it should not come as a surprise that with many long blocks, most of them if changed into iid lead to an increase of the LCS. To make this argument rigorous there are two problems we have to overcome:

1. Our result for one long block assumes that both sequences X and Y have length exactly equal to $2d$. An optimal alignment $\vec{a}$ of sequences of length n, will however "map" the pieces

$$X_1X_2X_3\dots X_{2d}\ ,\ X_{2d+1}X_{2d+2}\dots X_{4d}\ ,\ X_{4d+1}X_{4d+2}\dots X_{6d}\ ,\dots,X_{2d(m-1)+1}\dots X_n$$

   to pieces of Y of various lengths.

2. If $\vec{a}$ is an optimal alignment and say it aligns the piece

$$X_{2(i-1)d+1}X_{2(i-1)d+2}\dots X_{2id},$$

   with

$$Y_{j_1}Y_{j_1+1}\dots Y_{j_2},$$

   then the distribution of $Y_{j_1}Y_{j_1+1}\dots Y_{j_2}$ is no longer iid but rather complicated and poorly understood. Our result for the one long block case assumes however the Y-string to be iid.

To solve these two problems, we need a new idea which is introduced next via an example: Take $d=3$ and $n=18$, and consider the three intervals:

$$[1,2d]=[1,6],\ \ [2d+1,4d]=[7,12],\ \ [4d+1,n]=[13,18]. \tag{2.4}$$

We are going to specify the intervals to which these intervals should get aligned. For example we could align the first with $[1,7]$, the second with $[8,11]$ and finally the third with $[12,18]$. Within those constraints, we align in such a way to get a maximum number of aligned letter pairs (as usual we only allow same letter-pairs to be aligned with each other. Thus, one cannot align a letter with a different one). This means that in our current example, we align a maximum number of letter pairs of $X_1X_2\dots X_6$ and $Y_1Y_2\dots Y_7$. Then we align a maximal number of letter pairs of $X_7X_8\dots X_{12}$ and $Y_8\dots Y_{11}$. Finally we align $X_{13}\dots X_{18}$ with $Y_{12}\dots Y_{18}$ so as to get a maximum number of aligned letter pairs. The maximum number of aligned letter pairs under these constraints is hence equal to

$$|LCS(X_1X_2\dots X_6;Y_1\dots Y_7)|+|LCS(X_7X_8\dots X_{12};Y_8\dots Y_{11})|+|LCS(X_{13}X_{14}\dots X_{18};Y_{12}\dots Y_{18})|.$$

Note that the three terms in the sum on the right side of the above equality are independent. Of course the alignment defined in this way is not necessarily an alignment corresponding to a LCS. Indeed, let $k=2$, $n=12$ and the sequences $x=101010111111$

and $y = 001010011110$. Then the alignment which aligns 101010 with 00 and 111111 with 1010011110 is given by:

x		1	0	1	0	1	0			1		1			1	1	1	1	
y			0		0					1	0	1	0	0	1	1	1	1	0

(2.5)

In fact, it corresponds to two alignments: first the alignment aligning $X_1X_2 \dots X_6$ with Y_1Y_2:

1	0	1	0	1	0
	0		0		

(2.6)

and second the alignment aligning $X_7X_8 \dots X_{12}$ with $Y_3Y_4 \dots Y_{12}$:

1		1			1	1	1	1	
1	0	1	0	0	1	1	1	1	0

(2.7)

The alignment (2.5) is obtained by "concatenating" the alignments (2.6) and (2.7). The "score" of the alignment (2.5) is the sum of the scores of the alignment (2.6) and (2.7). Here, the alignment (2.6) aligned two letter pairs while (2.7) aligned six. Hence, when $X = x$ and $Y = y$, the score of the alignment (2.5) is

$$|LCS(x_1x_2 \dots x_6; y_1y_2)| + |LCS(x_7x_8 \dots x_{12}; y_3y_4 \dots y_{12})| = 2 + 6 = 8.$$

Now, the alignment corresponding to the LCS is

x			1	0	1	0	1	0		1	1	1	1		1	1
y		0		0	1	0	1	0	0	1	1	1	1	0		

(2.8)

and $LC_n = 9$. In the alignment (2.8) we have that $x_1x_2 \dots x_6$ gets aligned with $y_1y_2 \dots y_6$.

More generally, let $n = 2dm$, and let $0 = r_0 < r_1 < r_2 < \cdots < r_{m-1} \leq r_m = n$ be integers. Then, we study the best alignment under the following m constraints:

1) $X_1X_2 \dots X_{2d}$ gets aligned with $Y_1Y_2 \dots Y_{r_1}$

2) $X_{2d+1}X_{2d+2} \dots X_{4d}$ gets aligned with $Y_{r_1+1}Y_{r_1+2} \dots Y_{r_2}$

$\vdots$

m) $X_{2d(m-1)+1}X_{2d(m-1)+2} \dots X_n$ gets aligned with $Y_{r_{m-1}+1}Y_{r_{m-1}+2} \dots Y_n$.

The score of the best alignment under the above m constraints, denoted by $LC_n(\vec{r}) = LC_n(r_0, r_1, r_2, \dots, r_m)$, is equal to:

$$\begin{aligned} LC_n(\vec{r}) &= LC_n(r_0, r_1, r_2, \dots, r_m) \\ &:= \sum_{i=0}^{m-1} |LCS(X_{2di+1}X_{2di+2} \cdots X_{2d(i+1)}; Y_{r_i+1}Y_{r_i+2} \cdots Y_{r_{i+1}})|. \end{aligned}$$

Let now

$$\begin{aligned}\widetilde{LC}_n(\vec{r}) &= \widetilde{LC}_n(r_0, r_1, r_2, \ldots, r_{m-1}, r_m) \\ &= \sum_{i=0}^{m-1} |LCS(\tilde{X}_{2di+1}\tilde{X}_{2di+2}\cdots\tilde{X}_{2d(i+1)}; Y_{r_i+1}Y_{r_i+2}\cdots Y_{r_{i+1}})|, \end{aligned} \tag{2.9}$$

denote the score when the sequence X is replaced by the sequence $\tilde{X}$. Let $\mathcal{R}^n$ be the set of all the partitions of the integer interval $[0, n]$ into m pieces:

$$\mathcal{R}^n := \left\{(r_0, r_1, \ldots, r_m) \in [0, n]^{m+1} : r_0 = 0 \leq r_1 \leq r_2 \leq \cdots \leq r_m = n\right\},$$

let γ_k^e be a constant independent of d such that

$$\gamma_k^e < \gamma_k^*,$$

and let $0 < q^e < 1$ be the unique real, which exists by concavity, such that

$$\gamma_k(q^e) = \gamma_k^e.$$

Let $\varepsilon > 0$ and let

$$\mathcal{R}^n(\varepsilon) \subset \mathcal{R}^n,$$

be the subset of those element of $\mathcal{R}^n$ which have more than a proportion $1 - 2\varepsilon p$ of the values $r_i - r_{i-1}$ in the interval

$$\left[2d\frac{1-q^e}{1+q^e}, 2d\frac{1+q^e}{1-q^e}\right]. \tag{2.10}$$

More precisely, $(r_0, r_1, \ldots, r_m) \in \mathcal{R}^n(\varepsilon)$ if and only if

$$\text{Card}\left\{i \in \{1, \ldots, m\} : r_i - r_{i-1} \notin \left[2d\frac{1-q^e}{1+q^e}, 2d\frac{1+q^e}{1-q^e}\right]\right\} \leq 2mp\varepsilon.$$

With these notations, we then proceed to prove that with high probability every optimal alignment is in $\mathcal{R}^n(\varepsilon)$.

At this stage the reader, might wonder about the significance of the interval (2.10). The answer is found when we consider two independent iid strings where one has length $2d$ and the other has any length not in the interval (2.10). Then, the expected length of the LCS of two such strings, is less or equal to γ_k^e times the average of the lengths of the two strings. (To understand why, recall that the function γ_k is concave.) We can now use this for an alignment $\vec{a}$ between X and Y. Since, $\gamma_k^e < \gamma_k^*$, we infer that if too many of the pieces $X_{2di+1}X_{2di+2}\ldots X_{2d(i+1)}$ are to be matched by $\vec{a}$ with a piece of Y having length outside (2.10), then the score of the alignment $\vec{a}$ would, with high probability, be below optimal. Hence, $\vec{a}$ would typically not correspond to an LCS. This argument is made rigorous in the proof of Lemma 2.4.

Denote by $K^n(\varepsilon)$ the event that every optimal alignment is in $\mathcal{R}^n(\varepsilon)$ in other words, $K^n(\varepsilon)$ holds if and only if

$$\forall \vec{r} \in \mathcal{R}^n, \text{ such that } LC_n(\vec{r}) = LC_n, \text{ we have } \vec{r} \in \mathcal{R}^n(\varepsilon).$$

Let

$$\Delta(\vec{r}) = \Delta(r_0, r_1, \ldots, r_m) := \widetilde{LC}_n(r_0, r_1, \ldots, r_m) - LC_n(r_0, r_1, r_2, \ldots, r_m). \tag{2.11}$$

After proving that $K^n(\varepsilon)$ has high probability we show that with high probability, every alignment of $\mathcal{R}^n(\varepsilon)$ has a strong conditional increase. To do so, let $Q^n(\varepsilon)$ denote the event that for every alignment of $\mathcal{R}^n(\varepsilon)$ the conditional increase due to replacing a long block by iid is at least $d^\beta(\kappa(1-2\varepsilon)-2\varepsilon)$. More precisely, let

$$Q^n(\vec{r}) = \left\{ \mathbb{E}\left(\Delta(\vec{r}) \mid X, Y\right) \geq d^\beta(\kappa(1-2\varepsilon)-2\varepsilon) \right\},$$

where $\kappa > 0$ is a constant independent of n and d which will be specified later, and let

$$Q^n(\varepsilon) := \bigcap_{\vec{r} \in \mathcal{R}^n(\varepsilon)} Q^n(\vec{r}).$$

We will also need an event to insure that there are enough long blocks. For this, let O^n be the event that there are at least $mp/2$ long blocks, i.e.,

$$O^n = \left\{ \sum_{i=1}^{m} Z_i \geq \frac{mp}{2} \right\}.$$

The event K^n and $Q^n(\varepsilon)$ together imply the desired expected increase. This is the content of the next lemma

Lemma 2.1 *On $Q^n(\varepsilon) \cap K^n(\varepsilon)$,*

$$\mathbb{E}\left(|LCS(\tilde{X}; Y)| - |LCS(X; Y)| \middle| X, Y \right) \geq d^\beta(\kappa(1-2\varepsilon)-2\varepsilon) - 2\varepsilon.$$

Proof. Let $\vec{a}$ be an optimal alignment of X and Y. If $K^n(\varepsilon)$ holds, then $\vec{a}$ is an alignment in the set $\mathcal{R}^n(\varepsilon)$, and by the very definition of $Q^n(\varepsilon)$,

$$\mathbb{E}\left(\Delta(\vec{a}) \mid X, Y\right) \geq d^\beta(\kappa(1-2\varepsilon)-2\varepsilon).$$

Therefore,

$$\mathbb{E}\left(|LCS(\tilde{X}; Y)| - |LCS(X; Y)| \middle| X, Y \right) \geq d^\beta(\kappa(1-2\varepsilon)-2\varepsilon). \tag{2.12}$$

■

Now the above increase needs to be strictly positive to be of any use. We will see that holding $\kappa > 0$, fixed, we can take $\varepsilon > 0$ as small as we want and the event $Q^n(\varepsilon)$ will still have almost full probability, as long as d is taken large but fixed.

The bias (2.12), holds when $K^n(\varepsilon)$ and $Q^n(\varepsilon)$ both hold, therefore

$$\mathbb{P}\left(\mathbb{E}\left(|LCS(\tilde{X},Y)| - |LCS(X,Y)|\Big| X,Y\right) < d^{\beta}(\kappa(1-2\varepsilon)-2\varepsilon\right) \tag{2.13}$$

$$\leq \mathbb{P}((K^n)^c(\varepsilon)) + \mathbb{P}((Q^n)^c(\varepsilon)) \tag{2.14}$$

$$\leq \mathbb{P}((K^n)^c(\varepsilon)) + \mathbb{P}((O^n)^c) + \mathbb{P}((Q^n)^c(\varepsilon)\cap O^n). \tag{2.15}$$

The purpose of the next three lemmas is to show that the events $\mathbb{P}((K^n)^c(\varepsilon))$, $\mathbb{P}((O^n)^c)$ and $\mathbb{P}((Q^n)^c(\varepsilon)\cap O^n)$ hold with small probability.

Lemma 2.2 *Let $k\in\mathbb{N}$, $k\geq 2$ and let $\gamma_k^e<\gamma_k^*$. Let $\varepsilon>0$. Let $0<p<1$. Let d be such that $(1+\ln 2d)/2d \leq (\gamma_k^*-\gamma_k^e)^2p^2\varepsilon^2)/32$. Then,*

$$\mathbb{P}((K^n)^c(\varepsilon)) \leq \exp\left(-\frac{n(\gamma_k^*-\gamma_k^e)^2p^2\varepsilon^2}{32}\right),$$

for all $n=2dm$, $m\in\mathbb{N}$.

Proof. Let $\vec{r}=(r_0,r_1,r_2,\ldots,r_m)$ be an alignment in $\mathcal{R}^n$. Let $LC_n^*(\vec{r})$ denote the alignment score when we align X^* with Y according to $\vec{r}$:

$$LC_n^*(\vec{r}) := \sum_{i=0}^{m-1} |LCS(X^*_{2di+1}X^*_{2di+2}\cdots X^*_{2d(i+1)}; Y_{r_i+1}Y_{r_i+2}\cdots Y_{r_{i+1}})|,$$

and let LC_n^* denote the score of the LCS when we align X^* with Y:

$$LC_n^* := |LCS(X_1^*X_2^*\ldots X_n^*; Y_1Y_2\ldots Y_n)|.$$

When the alignment $\vec{r}$ does not belong to $\mathcal{R}^n(\varepsilon)$, then for n large enough, and as explained at the end of the present proof,

$$\mathbb{E}(LC_n^*(\vec{r}) - LC_n^*) \leq -\frac{3}{4}(\gamma_k^*-\gamma_k^e)p\varepsilon n. \tag{2.16}$$

Next, recall that $LC_n(\vec{r})$ denotes the alignment score when we align X with Y according to $\vec{r}$, while $LC_n := |LCS(X;Y)|$. The difference between X^* and X is at most m long block of length d^{β}. Hence the absolute difference between $LC_n^*(\vec{r})$ and $LC_n(\vec{r})$ is at most md^{β} and so is the absolute difference $|LC_n^*-LC_n|$. Therefore,

$$|(LC_n^*(\vec{r})-LC_n^*) - (LC_n(\vec{r})-LC_n)| \leq 2d^{\beta}m. \tag{2.17}$$

But if

$$LC_n(\vec{r}) - LC_n \geq 0, \tag{2.18}$$

is to hold, then, by (2.17), necessarily

$$LC_n^*(\vec{r}) - LC_n^* \geq -2d^{\beta}m. \tag{2.19}$$

Next, recall that $\beta < 1$, and choose d large enough so that

$$4d^{\beta} \leq (\gamma_k^* - \gamma_k^e) p\varepsilon d. \tag{2.20}$$

Combining (2.19) with (2.16) and (2.20) leads to:

$$LC_n^*(\vec{r}) - LC_n^* - \mathbb{E}(LC_n^*(\vec{r}) - LC_n^*) \geq \frac{1}{2}(\gamma_k^* - \gamma_k^e) p\varepsilon n, \tag{2.21}$$

and therefore

$$\mathbb{P}(LC_n(\vec{r}) - LC_n \geq 0) \leq \mathbb{P}\left(LC_n^*(\vec{r}) - LC_n^* - \mathbb{E}(LC_n^*(\vec{r}) - LC_n^*) \geq \frac{1}{2}(\gamma_k^* - \gamma_k^e) p\varepsilon n \right). \tag{2.22}$$

Now, $LC_n^*(\vec{r}) - LC_n^*$ depends on the iid random variables $X_1^*, X_2^*, \ldots, X_n^*$ and $Y_1, Y_2, \ldots, Y_n$, and changes by at most 2 when one of these variables changes. Therefore by Hoeffding's exponential inequality,

$$\mathbb{P}\left(LC_n^*(\vec{r}) - LC_n^* - \mathbb{E}(LC_n^*(\vec{r}) - LC_n^*) \geq \frac{1}{2}(\gamma_k^* - \gamma_k^e) p\varepsilon n \right) \leq \exp\left(-\frac{n(\gamma_k^* - \gamma_k^e)^2 p^2 \varepsilon^2}{16} \right). \tag{2.23}$$

Note that for $(K^n)^c(\varepsilon)$ to hold, we need at least one optimal alignment $\vec{r} \in \mathcal{R}^n$ which is not in $\mathcal{R}^n(\varepsilon)$. But if $\vec{r}$ is optimal then it corresponds to a LCS and thus $LC_n(\vec{r}) - LC_n \geq 0$. Hence,

$$(K^n)^c(\varepsilon) = \bigcup_{\vec{r} \notin \mathcal{R}^n(\varepsilon)} \{ LC_n(\vec{r}) - LC_n \geq 0 \},$$

so that

$$\mathbb{P}((K^n)^c(\varepsilon)) \leq \sum_{\vec{r} \notin \mathcal{R}^n(\varepsilon)} \mathbb{P}(LC_n(\vec{r}) - LC_n \geq 0).$$

The last sum above contains at most $\binom{n}{m}$ terms and so with the help of (2.23), we find

$$\begin{aligned}
\mathbb{P}((K^n)^c(\varepsilon)) &\leq \binom{n}{m} \exp\left(-\frac{n(\gamma_k^* - \gamma_k^e)^2 \varepsilon^2}{16} \right) \\
&\leq \left(\frac{ne}{m} \right)^m \exp\left(-\frac{n(\gamma_k^* - \gamma_k^e)^2 \varepsilon^2}{16} \right) \\
&\leq \exp\left(-\frac{n((\gamma_k^* - \gamma_k^e)^2 p^2 \varepsilon^2)}{16} + \frac{(1 + \ln 2d)}{2d} \right),
\end{aligned} \tag{2.24}$$

since $n = 2dm$. By our choice of d, $(1 + \ln 2d)/2d \leq (\gamma_k^* - \gamma_k^e)^2 p^2 \varepsilon^2 / 32$, leading to

$$\mathbb{P}((K^n)^c(\varepsilon)) \leq \exp\left(-\frac{n(\gamma_k^* - \gamma_k^e)^2 p^2 \varepsilon^2}{32} \right).$$

Let us next detail how the inequality (2.16) is obtained. This inequality only holds for alignments $\vec{r}$ which are not in $\mathcal{R}^n(\varepsilon)$. Hence, assume that $\vec{r} = (r_0, r_1, \ldots, r_m) \in \mathcal{R}^n$, but that $\vec{r} \notin \mathcal{R}^n(\varepsilon)$. Then, there are at least $2mp\varepsilon$ of the substrings

$$Y_{r_i+1} Y_{r_i+2} \ldots Y_{r_{i+1}}, \tag{2.25}$$

having their length not in the interval (2.10). If the string (2.25) has its length outside (2.10), then, as explained next, the expected value with the corresponding piece of X^* is at most γ_k^e times half the number of symbols involved. Thus, if the length of (2.25) is not in (2.10), then

$$\mathbb{E}|LCS(X^*_{2di+1}X^*_{2di+2}\dots X^*_{2d(i+1)}; Y_{r_i+1}Y_{r_i+2}\dots Y_{r_{i+1}})| \le \frac{\gamma_k^e}{2}(2d + r_{i+1} - r_i) \tag{2.26}$$

(To obtain the last inequality, use the fact that the expectation on the left side of (2.26) is by definition of $\gamma_k(\cdot,\cdot)$ (see (1.4)) equal to $\gamma_k(j, q^*)/2$ times the number of symbols j involved, where $q^* = (r_{i+1} - r_i - 2d)/j$, while $j = 2d + r_{i+1} - r_i$. Moreover, the function $t \to \gamma_k(t, q^*)$ is subadditive so that

$$\gamma_k(t, q^*) \le \gamma_k(q^*), \tag{2.27}$$

for all $t \in \mathbb{N}$. When $r_{i+1} - r_i$ is outside the interval (2.10), then q^* is outside of $[-q^e, q^e]$. But since the function γ_k is symmetric around the origin and concave,

$$\gamma_k(q^*) \le \gamma_k(q^e) = \gamma_k^e. \tag{2.28}$$

Combining (2.27) and (2.28), leads to

$$\gamma_k(j, q^*) \le \gamma_k^e.$$

This last inequality and the fact that the left side of (2.26) is equal to $\gamma_k(j, q^*)/2$ times the number of symbols involved, jointly imply the inequality (2.26).)

We can now apply a very similar argument for those i's, for which $r_{i+1} - r_i$ is in (2.10). For those i's, instead of inequality (2.26), we find the following:

$$\mathbb{E}\left|LCS(X^*_{2di+1}X^*_{2di+2}\dots X^*_{2d(i+1)}; Y_{r_i+1}Y_{r_i+2}\dots Y_{r_{i+1}})\right| \le \frac{\gamma_k^*}{2}(2d + r_{i+1} - r_i). \tag{2.29}$$

Combining (2.29) and (2.26), we have:

$$\begin{aligned}
\mathbb{E}LC_n(\vec{r}) &= \sum_{i=0}^{m-1} \mathbb{E}\left|LCS(X^*_{2di+1}X^*_{2di+2}\cdots X^*_{2d(i+1)}; Y_{r_i+1}Y_{r_i+2}\cdots Y_{r_{i+1}})\right| \\
&\le \frac{\gamma_k^e}{2} \sum_{\substack{i=0 \\ \vec{r}\notin\mathcal{R}_n(\varepsilon)}}^{m-1} (2d + r_{i+1} - r_i) + \frac{\gamma_k^*}{2} \sum_{\substack{i=0 \\ \vec{r}\notin\mathcal{R}_n(\varepsilon)}}^{m-1} (2d + r_{i+1} - r_i) \\
&= \frac{\gamma_k^*}{2} \sum_{i=0}^{m-1} (2d + r_{i+1} - r_i) + \left(\frac{\gamma_k^e - \gamma_k^*}{2}\right) \sum_{\substack{i=0 \\ \vec{r}\notin\mathcal{R}_n(\varepsilon)}}^{m-1} (2d + r_{i+1} - r_i) \\
&\le \frac{\gamma_k^*}{2}(2dm + n) + \left(\frac{\gamma_k^e - \gamma_k^*}{2}\right) \sum_{\substack{i=0 \\ \vec{r}\notin\mathcal{R}_n(\varepsilon)}}^{m-1} 2d \\
&\le \gamma_k^* n + \left(\frac{\gamma_k^e - \gamma_k^*}{2}\right) 2d\, 2mp\varepsilon \\
&= \gamma_k^* n - (\gamma_k^* - \gamma_k^e) np\varepsilon.
\end{aligned} \tag{2.30}$$

Next, as $n \to \infty$, $\mathbb{E}LC_n^*/n \to \gamma_k^*$. So, taking n large enough, we obtain that

$$\mathbb{E}LC_n^* \geq \gamma_k^* n - \frac{n}{4}(\gamma_k^* - \gamma_k^e)p\varepsilon. \tag{2.31}$$

In fact, our conditions on d, imply that (2.31) is satisfied by (1.2) for all $n = 2dm$. Combining now (2.30) and (2.31), gives the desired inequality (2.16). ∎

Lemma 2.3

$$\mathbb{P}(O^n) \geq 1 - \exp\left(-\frac{np^2}{4d}\right).$$

Proof. The total number of long blocks $\sum_{i=1}^m Z_i$ is a binomial random variable with parameters $m = n/2d$ and p. Thus,

$$1 - \mathbb{P}(O^n) = \mathbb{P}\left(\sum_{i=1}^m Z_i \leq \frac{mp}{2}\right) = \mathbb{P}\left(\sum_{i=1}^m Z_i - \mathbb{E}\sum_{i=1}^m Z_i \leq -\frac{mp}{2}\right) \leq \exp\left(-\frac{mp^2}{2}\right),$$

by Hoeffding's inequality. ∎

Lemma 2.4 *Let $\varepsilon > 0$. Let $0 < p < 1$. Let $\eta = 2\beta_1 - 1$, where $1/2 < \beta_1 < \beta < 1$. For d large enough,*

$$\mathbb{P}((Q^n)^c(\varepsilon) \cap O^n) \leq \exp\left(-Cmd^\eta \frac{p\varepsilon}{4}\right) = \exp\left(-Cn\frac{p\varepsilon}{8d^{1-\eta}}\right),$$

for all $n = 2dm$ and where $C > 0$ is a constant independent of d, n and ε.

Proof. We have already shown in [3] in the one long block situation, that changing the long block into iid tends to increase the LCS-score. For this check out Theorems 2.2 and 2.3 in [3] and the events O^d and H^d there. Now, these results are for the case when the two strings have length exactly equal to $2d$. However the same order of magnitude for the probability holds true, if the sequence Y has length close to $2d$ instead of exactly to $2d$, where by close to $2d$, we mean length in the interval (2.10). Thus, we have for all d that: for all j contained in the interval (2.10), we have

$$\begin{aligned}\mathbb{P}(|LCS(X_1^* \dots X_{2d}^*; Y_1 \dots Y_j)| - |LCS(X_1 \dots X_{2d}; Y_1 \dots Y_j)| \geq \kappa d^\beta | Z_1 = 1) \\ \geq 1 - \exp(-Cd^{2\beta_1 - 1}),\end{aligned} \tag{2.32}$$

where $\kappa > 0$ and $C > 0$ are constants not depending on d or j, but depending on the size of the alphabet, i.e., k. The above is obtained, for $j = 2d$, in the proof of Theorems 2.2 and 2.3 in [3]. The same proof holds true for j in the interval (2.10), so we leave it to the reader.

Now, let $\vec{r}$ be an alignment of $\mathcal{R}^n(\varepsilon)$, and let $M_\varepsilon^n(\vec{r})$ be the event that among the integers $i = 1, 2, \dots, m$, there are less than $mp\varepsilon/2$ of them for which: the length $r_i - r_{i-1}$ is in the interval (2.10) and there is a long block, but the LCS of $X_{2d(i-1)+1}X_{2d(i-1)+2} \dots X_{2di}$ with

$Y_{r_{i-1}+1}Y_{r_{i-1}+2}\dots Y_{r_i}$ does not increase by at least κd^β when we replace the long block by iid. Hence, $M_\varepsilon^n(\vec{r})$ is the event that the set

$$\left\{i \in \{1,2,\dots,m\} : Z_i = 1, r_i - r_{i-1} \in \left[2d\frac{1-q^c}{1+q^c}, 2d\frac{1+q^c}{1-q^c}\right], \Delta(\vec{r})_i < \kappa d^\beta\right\}$$

contains less than $mp\varepsilon/2$ elements. Here,

$$\Delta(\vec{r})_i := |LCS(X^*_{2d(i-1)+1}\dots X^*_{2di}; Y_{r_{i-1}+1}\dots Y_{r_{i+1}})| - |LCS(X_{2d(i-1)+1}\dots X_{2di}; Y_{r_{i-1}+1}\dots Y_{r_{i+1}})|.$$

Now, the probability for one such substring pair to have its LCS-score not increase by κd^β, is less than $\exp(-Cd^{2\beta_1-1})$, when $r_i - r_{i-1}$ is in (2.10) (by inequality (2.32)). This then leads to:

$$\mathbb{P}((M_\varepsilon^n)^c(\vec{r})) \le \binom{m}{\frac{mp\varepsilon}{2}} \exp\left(-Cd^{2\beta_1-1}\right). \tag{2.33}$$

Now

$$\binom{m}{\frac{mp\varepsilon}{2}} \le \left(\frac{2me}{mp\varepsilon}\right)^{\frac{mp\varepsilon}{2}} = e^{\frac{mp\varepsilon}{2}\left(1+\ln\frac{2}{p\varepsilon}\right)},$$

and therefore (2.33) becomes

$$\mathbb{P}((M_\varepsilon^n)^c(\vec{r})) \le e^{\frac{mp\varepsilon}{2}\left(1+\ln\frac{2}{p\varepsilon} - Cd^{2\beta_1-1}\right)}. \tag{2.34}$$

Note that the bound on the right side of the last inequality above is exponentially small in n when d is held fixed while n goes to infinity. (We also need d^γ to be larger than $2/(Cp\varepsilon\log_2 e)$.) Next note that when $M_\varepsilon^n(\vec{r})$, and O^n both hold, and $\vec{r} \in \mathcal{R}^n(\varepsilon)$, then

$$\mathbb{E}(\Delta(\vec{r}) \mid X, Y) \ge d^\beta(\kappa(1-2\varepsilon) - 2\varepsilon), \tag{2.35}$$

where, as defined before, $\Delta(\vec{r})$ is the change in score of the alignment $\vec{r}$, due to the modification of a randomly chosen long block into iid. (See (2.11).) The reason for the last inequality above is as follows: due to $M^n(\varepsilon)$, there are no more than $mp\varepsilon/2$ strings $X_{2d(i-1)+1}X_{2d(i-1)+2}\dots X_{2di}$ for which a change of the long block into iid does not create an increase of at least κd^β and for which $r_i - r_{i-1}$ is in the interval (2.10). But by O^n there are more than $mp/2$ long blocks. Each long block has the same probability to get chosen and changed to iid. Therefore the probability, to choose a long block so that the change does not create the increase we want but so that the corresponding $r_i - r_{i-1}$ has length in (2.10), must be less than $(mp\varepsilon/2)/(mp/2) = \varepsilon$. Similarly, since $\vec{r} \in \mathcal{R}^n(\varepsilon)$, there are no more than $mp\varepsilon/2$ of the integers $i \in \{1,2,\dots,m\}$, for which $r_i - r_{i-1}$ is not in the interval (2.10). So, the probability to chose a long block in $X_{2d(i-1)+1}X_{2d(i-1)+2}\dots X_{2di}$ with the corresponding $r_i - r_{i-1}$ outside (2.10), is also less or equal to ε. Next, we say that the string $X_{2d(i-1)+1}X_{2d(i-1)+2}\dots X_{2di}$ is with "defect" when either changing the long block does not create an increase of at least κd^β or that $r_i - r_{i-1}$ is not in (2.10). Let $T^n(\vec{r})$ be the event that the long block chosen is not with defect. From our discussion above, we find that when $O^n(\varepsilon)$ and $M_\varepsilon^n(\vec{r})$ both hold and $\vec{r} \in \mathcal{R}^n(\varepsilon)$, then

$$\mathbb{P}((T^n)^c(\vec{r}) \mid X, Y) \le 2\varepsilon. \tag{2.36}$$

Since

$$\mathbb{E}(\Delta(\vec{r}) \mid X, Y) = \mathbb{E}(\Delta(\vec{r}) \mid X, Y, (T^n)^c)\mathbb{P}((T^n)^c \mid X, Y) + \mathbb{E}(\Delta(\vec{r}) \mid X, Y, T^n)\mathbb{P}(T^n \mid X, Y), \tag{2.37}$$

and, since by its very definition, when T^n holds $\Delta(\vec{r})$ is greater or equal to $d^\beta\kappa$, it follows that,

$$\mathbb{E}(\Delta(\vec{r}) \mid X, Y, T^n) \geq d^\beta\kappa. \tag{2.38}$$

Moreover, since only one block of length d^β is changed, the change in score cannot be below $-d^\beta$, and therefore

$$\mathbb{E}(\Delta(\vec{r}) \mid X, Y, T^{nc}) \geq -d^\beta.$$

This last inequality combined with (2.38) and (2.36) in (2.37) yields,

$$\mathbb{E}(\Delta(\vec{r}) \mid X, Y) \geq -d^\beta 2\varepsilon + \kappa d^\beta(1 - 2\varepsilon) = d^\beta(\kappa(1 - 2\varepsilon) - 2\varepsilon). \tag{2.39}$$

Now (2.39) is obtained, assuming that $\vec{r} \in \mathcal{R}(\varepsilon)$ and that $M^n_\varepsilon(\vec{r})$ and $O^n(\varepsilon)$ both hold. By definition, the event $Q^n(\vec{r})$ is equivalent to inequality (2.39) and so for $\vec{r} \in \mathcal{R}^n(\varepsilon)$,

$$O^n \cap M^n_\varepsilon(\vec{r}) \subset Q^n(\vec{r}),$$

and thus

$$\mathbb{P}(O^n \cap (Q^n)^c(\vec{r})) \leq \mathbb{P}((M^n)^c_\varepsilon(\vec{r})). \tag{2.40}$$

Since

$$Q^n = \bigcap_{\vec{r} \in \mathcal{R}^n(\varepsilon)} Q^n(\vec{r}),$$

using (2.40), we get

$$\mathbb{P}(O^n \cap (Q^n)^c) \leq \sum_{\vec{r} \in \mathcal{R}^n(\varepsilon)} \mathbb{P}\left(M^n_\varepsilon(\vec{r})\right).$$

Applying (2.34) then gives

$$\mathbb{P}(O^n \cap (Q^n)^c) \leq \sum_{\vec{r} \in \mathcal{R}^n(\varepsilon)} e^{\frac{mp\varepsilon}{2}(1+\ln\frac{2}{p\varepsilon} - Cd^{2\beta_1-1})} \leq \binom{n}{m} e^{m(1-C\frac{p\varepsilon}{2}d^{2\beta_1-1})} \tag{2.41}$$

$$\leq \left(\frac{en}{m}\right)^m e^{m(1-C\frac{p\varepsilon}{2}d^{2\beta_1-1})} \tag{2.42}$$

$$= e^{-m(-2-2\ln 2d + C\frac{p\varepsilon}{2}d^{2\beta_1-1})} \tag{2.43}$$

Taking d such that

$$d^\eta > \frac{8 + 8\ln 2d}{Cp\varepsilon},$$

and combining it with (2.43), finally leads to

$$\mathbb{P}(O^n \cap (Q^n)^c) \leq \exp\left(-C\frac{p\varepsilon}{4}md^\eta\right),$$

which finishes this proof. ∎

Proof of Theorem 2.1. By Theorem 2.2, in order to prove that $\operatorname{Var} LC_n = \Theta(n)$ it is enough to show the high probability of a bias like in (2.1). However, the inequality (2.13) asserts that the probability of the bias

$$\mathbb{P}\left(\mathbb{E}\left(|LCS(\tilde{X};Y)| - |LCS(X;Y)|\big|X,Y\right) \leq d^{\beta}(\kappa(1-2\varepsilon)-2\varepsilon)\right), \tag{2.44}$$

is bounded above by

$$\mathbb{P}((K^n)^c(\varepsilon)) + \mathbb{P}((O^n)^c) + \mathbb{P}((Q^n)^c(\varepsilon) \cap O^n). \tag{2.45}$$

Lemma 2.2, 2.3 and 2.4, imply that the bound (2.45), is exponentially small in n. So, the expected change (2.44), is larger than $d^{\beta}(\kappa(1-2\varepsilon)-2\varepsilon)$ with probability close to one. The "close to one," is up to an exponentially small quantity in n. In order to apply Theorem 2.2, we would need a bias larger than $c_1 d^{\beta}$ where $c_1 > 0$ can be any constant not depending on n and d. To achieve this, simply take $\varepsilon > 0$ small enough so that

$$\kappa(1-2\varepsilon) - 2\varepsilon > 0. \tag{2.46}$$

e.g., $0 < \varepsilon = \kappa/4(\kappa+1)$. With this choice of ε, the bound (2.46) is equal to $\kappa/2$ where $\kappa > 0$ does not depend on n or d and, in turn, the expected conditional increase (2.44) is at least equal to $\kappa d^{\beta}/2$ with high probability. Hence, by Theorem 2.2, it follows that $\operatorname{Var} LC_n = \Theta(n)$, for d large enough but fixed. This finishes the proof. ∎

References

[1] K. S. Alexander. The rate of convergence of the mean length of the longest common subsequence. *Ann. Appl. Probab.*, 4(4):1074–1082, 1994.

[2] K. S. Alexander. Approximation of subadditive functions and convergence rates in limiting-shape results. *Ann. Probab.*, 25(1):30–55, 1997.

[3] S. Amsalu, C. Houdré, and H. Matzinger. Sparse long blocks and the microstructure of the LCS. ArXiv # math.PR/1204.49633, 2012.

[4] R. A. Baeza-Yates, R. Gavaldà, G. Navarro, and R. Scheihing. Bounding the expected length of longest common subsequences and forests. *Theory Comput. Syst.*, 32(4):435–452, 1999.

[5] J. Baik, P. Deift, and K. Johansson. On the distribution of the length of the longest increasing subsequence of random permutations. *J. Amer. Math. Soc.*, 12(4):1119–1178, 1999.

[6] F. Bonetto and H. Matzinger, Fluctuations of the longest common subsequence in the asymmetric case of 2- and 3-letter alphabets, *Alea*, 2:195–216, 2006.

[7] R. Bundschuh. High precision simulations of the longest common subsequence problem. *Eur Physical Journal B*, 22:533–541, 2001.

[8] R. Capocelli, A. De Santis, and U. Vaccaro, editors. *Sequences. II.* Springer-Verlag, New York, 1993. Methods in communication, security, and computer science, Papers from the workshop held in Positano, June 17–21, 1991.

[9] V. Chvàtal and D. Sankoff. Longest common subsequences of two random sequences. *J. Appl. Probability*, 12:306–315, 1975.

[10] V. Dančík and M. Paterson. Upper bounds for the expected length of a longest common subsequence of two binary sequences. *Random Structures Algorithms*, 6(4):449–458, 1995.

[11] J. Boutet de Monvel. Extensive simulations for longest common subsequences - Finite size scaling, a cavity solution, and configuration space properties. *Eur. Phys. J. B.*, 7:293–308, 1999.

[12] J. Boutet de Monvel. Mean-field approximations to the longest common subsequence problem. *Physical Rev. E*, 62:204–209, 2000.

[13] J. G. Deken. Some limit results for longest common subsequences. *Discrete Math.*, 26(1):17–31, 1979.

[14] C. Houdré, J. Lember, and H. Matzinger. On the longest common increasing binary subsequence. *C.R. Acad. Sci. Paris*, Ser. I 343:589–594, 2006.

[15] C. Houdré and H. Matzinger. On the variance of the optimal alignment-score for an asymmetric scoring function. ArXiv #math.PR/0702036 (forthcoming revision), 2007.

[16] A. R. Its, C. Tracy, and H. Widom. Random words, Toeplitz determinants, and integrable systems. I. Random matrix models and their applications. *Math. Sci. Res. Inst. Publ.*, 40, 2001.

[17] A. R. Its, C. Tracy, and H. Widom. Random words, Toeplitz determinants and integrable systems. II. Advances in nonlinear mathematics and science. *Phys. D.* 152–153:199–224.

[18] J. Lember and H. Matzinger. Standard deviation of the longest common subsequence. *Annals of Probability*, 37(3):1192–1235, 2009.

[19] S. N. Majumdar, K. Mallick, and S. Nechaev. Bethe ansatz in the Bernoulli matching model of random sequence alignment. *Phys. Rev. E*, 77:011110, 2008.

[20] S. N. Majumdar and S. Nechaev. Exact asymptotic results for the Bernoulli matching model sequence alignment. *Phys. Rev. E - Statistical Nonlinear and Soft Matter Physics*, 72:020901, 2008.

[21] M. J. Steele. Long common subsequences and the proximity of two random strings. *SIAM Journal on Applied Mathematics*, 42:731–737, 1982.

[22] M. J. Steele. An Efron-Stein inequality for non-symmetric statistics. *Annals of Statistics*, 14:753–758, 1986.

[23] C. Tracy and H. Widom. On the distributions of the lengths of the longest monotone subsequences in random words. *Prob. Theory Related Fields*, 119:350–380, 2001.

[24] M. S. Waterman. *Introduction to Computational Biology.* Chapman & Hall, 1995.

[25] M. S. Waterman. General methods of sequence comparison. *Bull. Math. Biol.*, 46(4):473–500, 1984.

[26] M. S. Waterman and R. A. Elton. Estimating statistical significance of sequence alignments. *Phil. Trans. R. Soc. Lond. B*, 344:383–390, 1994.